Text Alignment from Bimodal Mathematical Expression Sources

Sofiane Medjkoune, Harold Mouchère, Christian Viard-Gaudin, Simon Petitrenaud

2014 14th International Conference on Frontiers in Handwriting Recognition

Text Alignment from Bimodal Mathematical Expression Sources

Sofiane MEDJKOUNE, Harold MOUCHERE
and Christian VIARD-GAUDIN
LUNAM University, University of Nantes
IRCCyN UMR CNRS 6597
Rue Christian Pauc BP 50609 44306, Nantes, France
firstname.lastname@univ-nantes.fr

Simon PETITRENAUD
LUNAM University, University of Le Mans
LIUM - EA 4023
Avenue Laënnec, 72085 LE MANS CEDEX 9,
Le Mans, France
simon.petit-renaud@lium.univ-lemans.fr

***Abstract*—In this paper we propose a new approach to merge mathematical expression recognition results coming from handwriting and speech modalities. Using a bimodal description of mathematical expressions allows taking advantage of the complementarities between both signals, and can disambiguate situations were a single modality would not be clear enough. To combine the signals coming from both modalities, we propose to represent them in the same space as a textual description. First, from the handwriting signal, we generate the Nbest mathematical expressions; each of them is next translated as different possible strings. From the audio signal, an automatic speech recognition system provides a transcript, which is also available as a string. A string comparison algorithm is achieved to select the best mathematical expressions. This bimodal system is evaluated on real bimodal data from the HAMEX dataset and the results are compared to a single modality (handwriting) based system.**

I. Introduction

Mathematical expression (ME) recognition problem is attracting more and more interest within the scientific community. This is mainly due to the usefulness of the mathematical language and the challenges that this kind of problems raises. A particular representation in two dimensions with many special symbols has been developed for centuries, to facilitate the way that humans communicate mathematics with each others. Even if this graphical representation greatly assists in the transmission of the information conveyed by the studied mathematical principle, the insertion of such bi-dimensional elements in electronic documents is difficult. In fact, the bi-dimensional nature of ME, combined with the huge number of elementary symbols which are involved in its writing, increase the difficulty of interacting with a computer using mathematics based on traditional interfaces (mouse/keyboard). To recognize mathematical expressions, the technological progress offers alternative interaction modes that are more natural for human beings. In particular, speech and handwriting are among the most common ones. These modalities are very complementary, especially for mathematical equation description. A very common case is the following: a lecturer is writing a ME on a classroom blackboard and is dictating it in the same time to prevent from any misinterpretation from the audience (note that the two signals are not necessary synchronized). An example of such ambiguities is given on figure 1. To perform an automatic interpretation of either one or the other of both signals, some difficulties are encountered. These latter are intrinsic to each modality and the complementarity we discussed above can be

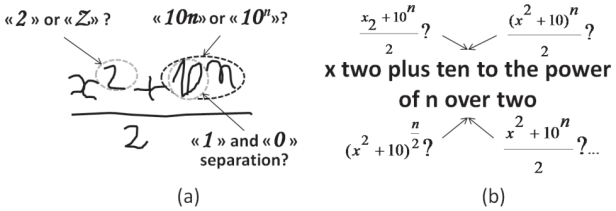

Fig. 1. Examples of intrinsinc ambiguities embeded by the (a) handwriting modality, (b) speech modality

used to increase the reliability of the interface in charge of mathematical expressions entering to a computer. Thus, in this work, we propose a bimodal architecture using the spoken and handwritten forms of the ME to recognize. More precisely, we exploit the Nbest list of mathematical expressions proposed by the system in charge of the handwriting modality, and by using a dedicated LaTeX string to text converter, we derive many different possible text translations. These translations are compared to the automatic transcription obtained from the system in charge of the speech signal interpretation. The best text alignment indicates the ME to keep as the best interpretation. The paper is organized as follows: in the second section, the handwriting based mathematical expression recognition (MER) is presented. Section III gives a short review on spoken MER. The fusion based MER approach we propose here is presented in section IV. We report the corresponding results in section V, and we conclude the paper in section VI.

II. Handwritten Mathematical Expression Recognition

We are considering online handwritten ME. This means that the raw data arriving to the handwriting recognition system is a sequence of elementary strokes which are ordered in time. In this work, we will consider that every symbol can be written with one or several strokes which are not necessarily consecutive, since some of them can be delayed. Most often, before starting the recognition process itself, the input signal undergoes a preprocessing step (spatial resampling, rescaling ...) [1], [2]. This preprocessing ensures consistency during the following processing steps, especially for the recognition one.

Generally, three sequential but interdependent steps have

2167-6445/14 $31.00 © 2014 IEEE
DOI 10.1109/ICFHR.2014.42

been identified to achieve handwritten ME recognition [1], [3]. The first step is the *segmentation* process in which the possible groups of strokes are formed. This stage is not trivial when it is supposed that interspersed symbols are authorized. Each *group* is called a segmentation hypothesis (*'sh'*). Ideally, each *'sh'* corresponds to a mathematical symbol. The recognition process is the second step. It aims to assign a symbol label (or a list of possible symbols) and a recognition score for each *'sh'*. The third step is the structural analysis. All the recognized symbols are used to make the final interpretation of the ME. This is done through a spatio-grammatical analysis. A drawback of such an approach, optimizing separately each step, is that the failure of one step can lead to the failure of the next one (error propagation). Rhee and Kim reported in [4] a solution to reduce this error propagation with the simultaneous optimization of the segmentation and recognition steps. However, in this case, the classifier is trained separately on isolated symbols. Later an improvement has been proposed by Awal and al. with a more global architecture [5]. The strengths of their system are the following. First of all, the recognition module is trained within the expressions directly from the outputs of the segmentation module. This allows a direct interaction between the different stages of the system (segmentation, recognition and 2D parsing). Secondly, during the segmentation step, a non-consecutive stroke grouping is allowed to form valid symbols. In addition, the classifier in charge of labeling each *'sh'* has the power to reject invalid hypotheses thanks to a *junk* class which is dedicated to label wrong segmentation hypotheses. Finally, the structural analysis (2D parsing) is controlled by both symbol recognition scores and a contextual analysis (spatial costs). The handwritten MER sub-part used in our architecture will be largely based on Awal and al.'s system.

III. Spoken mathematical expression recognition

Mathematical expression recognition based on automatic speech recognition (ASR) involves two main modules [6], [7]. The first one achieves the automatic speech recognition task. The output of this module provides a textual description which depends of the audio description and of the ASR reliability. This text is composed of words written with alphabetic characters as they are recognized by the ASR system. This text is ideally a fair description of the ME (it depends on the quality of word pronunciation by the speaker). Fig. I (b) gives an example of a possible recognized string by the ASR system "x two plus ten to the power of n over two". The second module is a parser, which processes the previous transcription in the 2D space to deduce the associated ME.

The automatic transcription is given by an ASR system which is quite similar to the one described in the case of handwriting modality. The main difference is the nature of the signal which is processed (acoustic in this case). This recognition procedure involves three stages. During the first one, the acoustic signal is filtered and re-sampled, then a frame description is produced, where a feature vector is computed for each window of 25 ms with an overlap of 10 ms. The features are the cepstral coefficients and their first and second derivatives [8]. Segmentation into homogeneous parts is operated in a second step. Resulting segments are close to minimal linguistic units. The last step is the decoding itself using models and tools learned within a training step (acoustical model, pronunciation dictionary and language model).

Parsing the resulting transcription from the previous module is a very hard task. In the rare existing systems [6], [7], the parsing is most of the time assisted by either introducing some dictation rules (in order to separate the numerator and the denominator of a fraction, for instance) or using an additional source of information (such as using a mouse to point the position where to place the different elements). By adding such constraints, the editing process becomes less natural and far from what is expected from this kind of systems.

The work we report in this paper concerns the French spoken language. The task of speech recognition in our system is carried out by a system largely based on the one developed at the LIUM [8], which kernel is one of the most popular worldwide speech recognition systems (CMU-Sphinx)[9].

IV. Bimodal mathematical expression recognition

A. The data fusion principle

The idea of multi-modal human-machine interaction comes from the observation of the human beings' interaction. Usually, people simultaneously use many communication modes to converse. In so doing, the conversation becomes less ambiguous. The main goal of this work is to mimic this procedure to be able to set up a multi-modal system dedicated to mathematical expressions recognition (MER).

Generally, data fusion methods are divided in three main categories [10], [11]: *early fusion* which happens at features levels; *late fusion* which concerns the intermediate decisions fusion and the last one is the *hybrid fusion* which is a mix of the two. Within each approach, three kinds of methods can be used to carry out the fusion process. Rules based approaches represent the first category and include methods using simple operators such as max, (weighted) mean or product. The second category is based on classifiers and the last one is based on parameter estimation.

Since we are interested in combining two heterogeneous signals (handwriting and audio streams), we decided to consider a late fusion strategy, to ensure to use suitable recognition systems with respect to each modality. In a previous work [12], we have merely proposed a bag of words approach to combine information coming from the audio description in the main stream processing of the handwritten signal. The problem of alignment of the two streams was not investigated during these previous works, which can highly penalize the combination process. Thus, the matter of this paper is to consider the audio and handwriting streams alignment in order to improve the global performance of the system. In the following sections we describe the architecture of the proposed collaborative system.

B. Data fusion for mathematical expression recognition

The proposed architecture for bimodal mathematical expressions recognition (BMER) is presented in Fig. 2. Its overall description is as follows.

The system input is a ME available at both spoken and handwritten forms. An automatic speech recognition (ASR)

system is in charge of the interpretation of the speech signal describing the ME (*cf.* section III). The output of this system is a text describing the ME. At this level, the result is still one dimensional, it is a standard text. This textual description is composed of two categories of words, the first one concerns words which are useful in a mathematical language point of view, we call them *keywords*. The other category includes all the other words which are present in the text, as *stop words*, only for linguistic fluentness. Only the keywords are of interest for us, thus we automatically filter the textual description to keep only this category of words. Similarly, the handwriting recognition module processes the on-line handwritten signal to form the basic symbol hypotheses from the raw signal (sequence of strokes), as explained in section II. After this step, a list of labels and their corresponding scores are assigned to each symbol. The set of resulting symbols is then parsed in the $2D$ plan to define the ME layout. In the previous work [12], we considered that the fusion process can be carried out at two levels: directly during the symbol recognition step and next during the structural analysis to identify the spatial relationships. In this novel approach we propose here, we delay the fusion operation until the interpretation step of the full ME by the handwriting based system. The main idea is to run the full process of ME recognition considering only the handwriting modality. This system is able to provide an $Nbest$ list of possible LaTeX strings corresponding to the input signal. Once this $Nbest$ list obtained, each LaTeX strings is sent to the *LaTeX2Text* module which elaborates various possible translations with regard to each LaTeX string. Accordingly, the audio stream, through its associated automatic transcription, is exploited to make a re-ranking of the list proposed by the handwriting system. This merging process is done with the help of the fusion units (grey boxes) in the architecture of figure 2 described below.

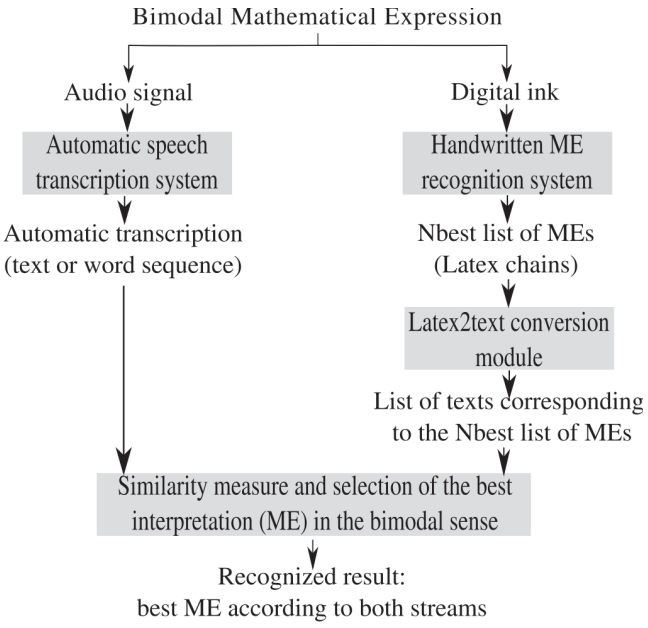

Fig. 2. The collaborative architecture for complete mathematical expression recognition

1) LaTeX2Text conversion module: The role of this module is to give a textual description associated with the LaTeX chain of a ME. This textual description is intended to be the most natural. It should be as close as possible to the description of a dictation provided by a speaker. Since various dictations are possible for a given ME LaTeX chain, the generator gives many different ways to dictate the same expression. In table I is given an example of such a procedure.

TABLE I. EXAMPLE OF DIFFERENT POSSIBLE TRANSLATIONS FOR A SAME LaTeX STRING

LaTeX chain	Associated translations in French (in English)
`$\frac{x^2+10^n}{2}$`	- *"x carré plus dix n sur deux"* (*x squared plus ten n over two*) - *"x au carré plus dix puissance n le tout sur deux"* (*x squared plus ten to the power n all over two*) - *"x puissance deux plus dix à la puissance n sur deux"* (*x to the power two plus ten to the power n over two*) - *"x carré plus dix puissance n divisé par deux"* (*x squared plus ten to the power n divided by two*) ⋮

2) Similarity measure and best solution selection module: The $Nbest$ list of MEs given by the handwriting recognition system and initially represented in a LaTeX form, are now represented in the textual description space. Each LaTeX chain produces one or more string(s). Each of these textual descriptions is compared to the transcription issued from the ASR system and a score is associated to each measure. The highest score gives the best matching and consequently the best ME to consider as the final solution. The similarity measure is based on the *Levenshtein distance* between strings. This metric is composed of three quantities: the number of substitutions denoted *Subst*, the number of the deletions (*Del*) and the number of the insertions (*Ins*). If this measure is normalized considering the number of words in the reference (the automatic transcription from the speech system in our case), denoted by *N*, we obtain the **W**ord **E**rror **R**ate (WER) metric. Equation 1 defines the WER which can exceed 100% since sometimes it is required to perform several operations to recover one word of the reference string:

$$WER = \frac{Ins + Del + Subst}{N} \quad (1)$$

It is also possible to compute the **W**ord **C**orrect **R**ate (WCR) as defined by equation 2, which is bounded between 0 and 100%:

$$WCR = 1 - \frac{Subst + Del}{N} \quad (2)$$

We select the solution in the $Nbest$ list which gives the best WCR with respect to the transcription of the audio signal. Indeed, since our goal is to keep the text describing a ME which has the higher common number of words compared to the audio description, we are not interested by the number of inserted words (Ins).

This similarity measure is calculated on preprocessed texts to remove all stop words and consider only keywords, as explained before.

V. Experimental results

In this section we present the performances of the system reported in this paper. First, we give an overview of the dataset we used. Then the performances of the mono-modal systems are presented. After that we report the results concerning our system compared to the baseline system based only on the handwriting signal and also compared to the previous architecture, we presented in [12] (based on fusion at lower levels: symbols and relations).

A. Dataset description

The data used to perform the experiment is from the *HAMEX* [13] database. This database includes a set of approximately 4 350 ME, each of them available in the spoken and the handwritten modalities. The vocabulary covered by *HAMEX* contains 74 mathematical symbols, including all the Latin alphabet letters, the ten digits, six letters from the Greek alphabet and various mathematical symbols (integral, summation...).

B. Specialized systems performance

The handwriting recognition task is accomplished with the on-line handwritten MER system that participated to $CROHME2012$[1] competition [14]. The results reported here concern a set of 519 MEs of the *HAMEX* test part which satisfies the *CROHME* (task 2) grammar and vocabulary (56 symbol classes). A set of 500 MEs of the *HAMEX* train part satisfying the same conditions as before are used to tune the different parameters we consider in the fusion system. Finally, the models of the ASR system are trained on the whole speech data of the HAMEX train part. Concerning the fusion process itself, the value of $Nbest$ ME is set experimentally to 10 using the validation database. We report on Table II the performances of the handwriting system considering that the valid solution is ranked first ($TOP1$), or ranked among the first two answers ($TOP2$) and so on.

TABLE II. Performances of the handwriting recognition system

Evaluation level	*TOP1*	*TOP2*	*TOP3*	*TOP4*	*TOP5*	*TOP10*	*more*
Reco. rate [%]	34.10	42.08	44.6	45.6	45.75	47.68	48.36

In another side, the recognition rate of the automatic speech transcription system with respect to the whole vocabulary guiding the ASR system is 90.06%. If only keywords are considered for the evaluation, the recognition rate is increased to 97.21%. This rate is given at the word level, not as in Table II at the expression level, since at that stage the interpretation of the ME is not yet achieved.

As we can observe from Table II, the handwriting modality based system gives the right interpretation of the input signal in 34.1% of cases. If the first two answers are considered, 8% more are saved and if ten best solutions are taken into account from the output of this system we reach a recognition rate of 47.68%.

[1] http://www.isical.ac.in/~crohme/index.html

This observation combined with the performance with respect to the speech modality suggests that the combination of both modalities should increase the $TOP1$ recognition rate obtained with the handwriting based system alone. In the following the results of this procedure are reported.

C. The proposed system performances

Table III reports the comparison of the handwritten mathematical expressions recognition system and the bimodal based one, considering the fusion at symbols and relations levels [12] and considering the approach proposed here.

TABLE III. Comparison of the performances of the handwriting recognition system with the fusion based system proposed in [12](Syst. I) and the one proposed here (Syst. II)

Recognition rate in [%] of	Strokes	Symbols	Expressions with		
			Exact match	1 error at most	2 errors at most
Handwriting based system	80.05	82.93	34.10	46.44	49.52
Syst. I	86.73	88.21	41.82	50.67	53.37
Syst. II	86.65	89.30	42.00	51.06	52.02

In Table III we can observe that the system we proposed here (Syst. II) outperforms significantly the baseline system based on handwriting signal and only slightly the multi-modal system we proposed in [12] where the fusion is achieved at symbols and relations levels. It is clear that the bimodal aspect of the information allows not only to improve the recognition at the expression level, but also at the lower levels (strokes and symbols).

With the proposed approach (Syst. II), every solution that is in the handwritten $Nbest$ list is treated equally with respect to the audio transcript. In another words, the last proposal of the list could be selected if its similarity measure is the best one with respect to the audio transcript, even if the initial cost of this solution is very high compared to the $TOP1$ solution. To prevent this situation, we propose a variant of the proposed method, using a reject threshold to possibly shorten the NBest list.

In this regard, the new $Nbest$ list, denoted $Nbest'$ is given by equation 3:

$$Nbest' = \{Topj \in Nbest \ / \ \frac{|C_1 - C_j|}{C_1} \leq \alpha\}, \qquad (3)$$

where α is a parameter we fixed experimentally to 1.3 using the validation database. The variables C_1 and C_j are the initial costs (given by the handwriting based system) associated with the $Top1$ and $Topj$ MEs respectively.

Therefore, every solution which has a relative cost more than alpha times the $Top1$ solution will be discarded from the list.

The obtained results with this system (Syst II') are reported in the Table IV. It shows that the use of restricted solutions with too low relative costs improves the performances at all levels (stroke, symbol and ME). Here, the interesting point is

that the gain in term of ME recognized with one or two errors is very significant compared to the previous systems (Syst. I and Syst. II).

TABLE IV. COMPARISON OF THE PERFORMANCES OF THE HANDWRITING RECOGNITION SYSTEM WITH THE FUSION BASED SYSTEM PROPOSED IN [12](SYST. I) AND THE EXTENSION OF THE ONE PROPOSED HERE (SYST. II')

Recognition rate in [%] of	Strokes	Symbols	Expressions with		
			Exact match	1 error at most	2 errors at most
Handwriting based system	80.05	82.93	34.10	46.44	49.52
Syst. I	86.73	88.21	41.82	50.67	53.37
Syst. II'	87.13	90.83	42.97	53.09	57.34

The improvement brought by the current method with respect to the previous bimodal system (Syst. I) is not necessarily important, however it gives another point of view of where the fusion can happen (at the ME interpretation level). In addition, there are more and more MEs which are recognized with one or two errors than compared to Syst. II and Syst. I.

VI. CONCLUSION AND FUTURE WORK

In this work we presented a new approach for bimodal mathematical expressions recognition. The modalities in concern are speech and handwriting.

The main novelty of this work is to consider the combination process during the interpretation step. This procedure allows to prevent from the problem of the existing asynchrony between both streams during processing at lower levels (symbols and elementary relations).

The reported results showed the interest of such a processing. This can be seen either at expression level and in lower levels (strokes and symbols).

In a future work, as a first extension of the Syst. II', we plan to use both handwriting costs and similarity measures in a global cost function to give the final interpretation. We are also planning to exploit the two strategies of fusion we investigated (Syst. I and Syst. II') in order to tend to a more complete system where the bimodal information is exploited during symbols/relations identification and during the interpretation.

REFERENCES

[1] B. Dorothea and G. Ann, *Recognition of mathematical notation*, H. Bunke, P. Wang ed. Handbook of Character Recognition and Document Image Analysis, 1997.

[2] E. Tapia and R. Rojas, "A survey on recognition of on-line handwritten mathematical notation," Free University of Berlin, Tech. Rep., 2007.

[3] K. F. Chan and D. Y. Yeung, "Mathematical expression recognition: A survey," *International Journal of Document Analysis and Recognition*, vol. 3(1), pp. 3–15, 2000.

[4] T. H. Rhee and J. H. Kim, "Robust recognition of handwritten mathematical expressions using search-based structure analysis," in *Proc. of Int. Conf. on Frontier in Handwriting Recognition (ICFHR)*, 2008, pp. 19 – 24.

[5] A.-M. Awal, H. Mouchère, and C. Viard-Gaudin, "A global learning approach for an online handwritten mathematical expression recognition system," *Pattern Recognition Letters*, no. 35, pp. 68–77, 2014.

[6] R. Fateman, "How can we speak math," *Journal of Symbolic Computation*, vol. 25, no. 2, 1998.

[7] A. Wigmore, G. Hunter, E. Pflugel, J. Denholm-Price, and V. Binelli, "Using automatic speech recognition to dictate mathematical expressions: The development of the talkmaths application at kingston university," *Journal of Computers in Mathematics and Science Teaching (JCMST)*, vol. 28(2), pp. 177–189, 2009.

[8] P. Deléglise, Y. Estève, S. Meignier, and T. Merlin, "Improvements to the lium French ASR system based on cmu sphinx: what helps to significantly reduce the word error rate?" in *Proc. of Int. Conf. Interspeech*, 2009, pp. 2123–2126.

[9] "Cmu sphinx system," http://cmusphinx.sourceforge.net, Accessed on February, 20th, 2014.

[10] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: A survey," *Multimedia Systems*, vol. 16(6), pp. 345–379, 2010.

[11] J. P. Thiran, F. Marqués, and H. Bourlard, *Multimodal Signal Processing - Theory and Applications for Human-Computer Interaction*. Elsevier, 2010.

[12] S. Medjkoune, H. Mouchere, S. Petitrenaud, and C. Viard-Gaudin, "Multimodal mathematical expressions recognition: Case of speech and handwriting," in *Human-Computer Interaction. Interaction Modalities and Techniques*, ser. Lecture Notes in Computer Science, M. Kurosu, Ed. Springer Berlin Heidelberg, 2013, vol. 8007, pp. 77–86.

[13] S. Quiniou, H. Mouchère, S. Peña Saldarriaga, C. Viard-Gaudin, E. Morin, S. Petitrenaud, and S. Medjkoune, "HAMEX - a handwritten and audio dataset of mathematical expressions," in *Proc. of Int. Conf. on Document Analysis and Recognition (ICDAR)*, 2011, pp. 452–456.

[14] H. Mouchère, C. Viard-Gaudin, D. H. Kim, J. H. Kim, and U. Garain, "ICFHR2012: Competition on recognition of online handwritten mathematical expressions (crohme 2012)," in *Proc. of Int. Conf. on Frontier in Handwriting Recognition (ICFHR)*, 2012, pp. 811–816.